Homework 10

Project W6: KAGGLE-YOUTUBE TRENDING VIDEOS
Laima Anna Dalbina, Eva Ibrus, Annika Jaakson

Task 2. Business understanding

Identifying business goals

Background

Youtube, one of the biggest video sharing platforms in the world, has a special section for videos that are “trending” or in other words, quickly growing in popularity. However, the fact that those videos are gaining popularity very fast doesn’t mean that they will gather a large number of views in the long term. Since the section contains up to 200 videos that change every day, then most of these videos “trend” for only one day in the bottom of the list. This means that the list contains many videos which aren’t actually that popular but just get a boost of views over a short period of time.

This “trending list” is also different for each country, meaning that videos gaining popularity in India, for example, might be completely unknown in other places of the world. When having data from multiple countries’ trending list, analysis on the differences between them can be made, thus creating a better understanding of the regional peculiarities of Youtube.

Business goals

Our goal is to find interesting data and trends about the videos. Are there identifiable common traits within video titles, tags or descriptions? Do the views for trending videos grow the same way? Can we predict how long a video stays on the trending page? These are only some of the questions we hope to answer with this project.

The information might be useful for people making videos and uploading them to Youtube, but since we can only provide insight on videos that are already trending, then it won’t help anybody get into the trending page.

Business success criteria

This section is not really relevant for our project, because our goal is to find interesting data and that can be subjective.

Assessing our situation

Inventory of resources

We have data from 10 different countries (Canada, Denmark, France, Great Britain, India, Japan, Korea, Mexico, Russian and USA). Each day there are up to 200 trending videos from 14.11.2017 till 31.05.2018.

The people working on this project are Annika Jaakson, Eva Ibrus and Laima Anna Dalbina. Mentors are Meelis Kull, Markus Kängsepp and Laura Ruusmann. We will be using the programming language Python 3 and its data science-related libraries (pandas, matplotlib, scipy and others) to analyse and visualise the data. We will conduct our work in the Jupyter Notebook programming environment. For collaboration we shall use Git and all of our code can be accessed at Github. No specific hardware is used for this project.

Requirements, assumptions, and constraints

The project must be completed by 16th December. By that time there should be a complete project on Github and a poster outlining the most interesting finds which will be presented on 19th December.

Risks and contingencies

The only factor that could delay the completion of the project is running out of time. This will be prevented by careful planning and dividing the tasks among project members. If one member feels that they are not able to complete their task or it will take too much time, they should immediately contact other members so the task could be taken over by someone else.

Terminology

Trending video - a video on the trending page of Youtube determined by an algorithm. Generally the following factors are considered: view count, how fast the video generates views, the age of the video etc.

Costs and benefits

Not relevant for our project - our dataset is freely available on the Internet and noone is getting paid.

Defining our data-mining goals

Data-mining goals

We need to produce a clean dataset which corresponds to our needs and which has correct formatting so that we could start analyzing it. The project should produce several graphs

showing the rate of views/comments/likes growth, how long the videos stay in trending, word maps with the most popular words in titles/tags/descriptions.
Our machine learning model should predict how quickly a video gets into the trending page and how long it stays there.

Data-mining success criteria

Not relevant for our project, as our goal is to find interesting trends and that can be subjective.

Task 3. Data understanding

Gathering data

Outline data requirements

The data needed to address the goals should consist of videos from at least 5 different countries. The data should be gathered every day consistently for at least 3 months to make any relevant analysis and predictions. Each file should consist of the particular video details. For example, the most crucial properties are channel title, video title, video id, trending date and publish time to even look at a trending video. To predict what influences being on the top of the trending page, properties like category, tags, views, likes, dislikes and comments are necessary. The data should preferably be in csv format to simplify the analysis part.

Verify data availability

Our dataset acquired from Kaggle (https://www.kaggle.com/datasnaek/youtube-new) is very suitable to meet our project goals. It consists of data from 10 different countries. The data has been gathered for 6 months and 18 days which fits our requirements. The dataset has all of the data necessary to address the requirements stated above. Additionally, there is a description for each of the videos.

Define selection criteria

We will use the dataset mentioned above. As the dataset does not have many properties, almost every one of them are useful for us. We will not use thumbnail_link because it is not needed for any of our goals. Everything else is relevant.

Describing data

The dataset consists of 10 csv files and 10 json files. Every separate file is meant for each country and there is a corresponding json file for each country which consists of the categories which are stated in the csv files as a number.
Every csv file consists of the exact same columns. There are 16 columns for each file. They are video_id, trending_date, title, channel_title, category_id, publish_time, tags, views, likes, dislikes, comment_count, thumbnail_link, comments_disabled, ratings_disabled, video_error_or_removed and description.
The csv files are organized by dates, meaning that for every day in the data gathering period (the specific dates vary somewhat between countries) there is one row for each video that was in the trending list for that day. This means that many videos appear several times in the dataset under

multiple days. However, there are some imperfections in the dataset, for example the Japan file contains data from only 122 days, not 205 like the others and some files contain considerably more unique videos than others.

Exploring data

While exploring the data we found some issues, mostly with empty values. In the video_id column there are several entries which are just #NAME (2312 out of 375942). Solution: replace these values with NaN or generate an id (we can find if it's the same video by checking the publish time and video title).

There are videos where the tags are [none] - can be replaced with NaN as well.

There are some videos which have been deleted at some point - for these the title is 'Deleted video' - these can be replaced with NaN as well.

The tags column is formatted weirdly - the tags are separated by | and are enclosed with "", except for the first and last tag.

There are two columns with datetime - trending_date and publish_time but they are formatted differently.

While checking the category json files we found that one category (Nonprofits & Activism) is missing in most of the json files, although the corresponding countries have videos in that category.

Verifying data quality

The chosen dataset meets our requirements and has the data we need. There are some issues with data quality but these can be corrected quickly.

Task 4. Planning our project

Task 0: Planning

Task assigned to: 6h for each person

Task 1: Cleaning the data

Task assigned to: Eva, estimated time 6h

Task 2: Most popular feature attributes (for all countries and for each country separately)

Task assigned to: Laima, estimated time: 3h

Task 3: Find how quickly do videos get into the trending page and how long does a video stay trending

Also see if those times are different for different countries, different categories and most popular channels.

Task assigned to: Annika, estimated time 5h

Task 4: Word Maps for tags, titles, descriptions

Task assigned to: Laima, estimated time: 5h

Task 5: Find how many videos are popular globally versus locally

For example, see how many trending videos in India are also trending in other countries (and which are trending only in India).

Task assigned to: Annika, estimated time: 4h

Task 6: Analyse differences between English-speaking countries

Since Canada, USA and UK all speak English, they should also have more trending videos in common than other countries (and also similar tags and words in titles etc). We want to see if this hypothesis is true.

Task assigned to: Annika, estimated time: 6h

Task 7: Analysing the views in detail

See how the views change. Is there a difference in the views when a video gets to trending page?

Task assigned to: Eva, estimated time 7h

Task 8: Look at the first half and the last half of the trending videos to see what it takes to get into the first half.

As there are approximately 200 videos from the trending tab every day recorded, we can look at the first half and the last half to check what are the differences between them.

Task assigned to: Laima, estimated time: 5h

Task 9: Implementing machine learning models

Predict how quickly a video gets into trending page, how many days it stays in the trending page (using regression models). We could also predict whether a video stays on trending for longer than 1 day.

Estimated time: 6h for each person

Task 10: Making the poster and presentation

Task assigned to: 6h for each person